Fusing Continuous-valued Medical Labels using a Bayesian Model

Tingting Zhu, Nic Dunkley, Joachim Behar, David A. Clifton,

Gari D. Clifford

arXiv:1503.06619v2 [cs.LG] 13 Jun 2015

The authors are with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, United Kingdom. G D Clifford is also with the departments of Biomedical Informatics and Biomedical Engineering at Emory University and Georgia Institute of Technology.

Contact for Correspondence: Tingting Zhu

E-mail: tingting.zhu@eng.ox.ac.uk

Abstract With the rapid increase in volume of time series medical data available through wearable devices, there is a need to employ automated algorithms to label data. Examples of labels include interventions, changes in activity (e.g. sleep) and changes in physiology (e.g. arrhythmias). However, automated algorithms tend to be unreliable resulting in lower quality care. Expert annotations are scarce, expensive, and prone to significant inter- and intra-observer variance. To address these problems, a Bayesian Continuous-valued Label Aggregator(BCLA) is proposed to provide a reliable estimation of label aggregation while accurately infer the precision and bias of each algorithm.

The BCLA was applied to QT interval (pro-arrhythmic indicator) estimation from the electrocardiogram using labels from the 2006 PhysioNet/Computing in Cardiology Challenge database. It was compared to the mean, median, and a previously proposed Expectation Maximization (EM) label aggregation approaches. While accurately predicting each labelling algorithm's bias and precision, the root-mean-square error of the BCLA was 11.78 ± 0.63ms, significantly outperforming the best Challenge entry (15.37 ± 2.13ms) as well as the EM, mean, and median voting strategies (14.76 ± 0.52ms, 17.61 ± 0.55ms, and 14.43 ± 0.57ms respectively with $p < 0.0001$).

The BCLA could therefore provide accurate estimation for medical continuous-valued label tasks in an unsupervised manner even when the ground truth is not available.

Keywords Crowdsourcing · Bayes methods · Time series analysis · Electrocardiography.

1 Introduction

With human annotation of data, significant intra- and inter-observer disagreements exist [7, 21]. Expert labelling (or 'reading' or 'annotating') of medical data by physicians or clinicians often involves multiple over-reads, particularly when an individual is under-confident of the diagnosis. However, experts are scarce and expensive and can create significant delays in labelling or diagnoses. Although medical training includes periodic assessment of general competency, specific assessments for reading medical data are difficult to be performed regularly. This data processing pipeline is further complicated by the ambiguous definition of an 'expert'. There is no empirical method for measuring level of expertise, even though label accuracy can vary greatly depending on the expert's

experience. As a result, there exists a great deal of inter- and intra-expert variability among physicians depending on their experiences and level of training [7, 13, 14, 17, 18, 21].

An effective probabilistic approach to aggregating expert labels which used an Expectation Maximization (EM) algorithm, was first proposed by Dawid and Skene [6]. They applied the EM algorithm to classify the unknown *true* states of health (i.e. fit to undergo a general anaesthetic) of 45 patients given the decision made by five anaesthetists. Raykar *et al.* [16] extended this approach to measure the diameter of a suspicious lesion on a medical image using a regression model. Their assumption was that the discrepancies of the lesion diameter estimates from different expert annotators were Gaussian distributed and noisy versions of the actual *true* diameter. The precision of each expert annotator and the underlying ground truth were jointly modelled in an iterative process using EM. Welinder and Perona [23] proposed a Bayesian EM framework for continuous-valued labels, which explicitly modelled the precision only of each annotator to account for their varying skill levels, without modelling the bias of annotators. A more specialised form of the Bayesian model of bias was proposed by Welinder *et al.* [22] but for binary classification tasks. However, their model cannot account for more complex tasks such as the continuous-valued labelling.

The methodology proposed in the work presented in this article improves on these prior algorithms [16, 22, 23] by introducing the novelty of combining *continuous-valued annotations* to infer the underlying ground truth, while *jointly modelling the annotator's bias and precision* in an unified model using a Bayesian treatment.

Aggregating annotations (i.e. fusing multiple annotations for each piece of data from annotators with varying levels of expertise) from human and/or automated algorithms may provide a more accurate ground truth and reduce annotator inter- and intra-variability. However, most annotators are likely to have some bias regardless of their expertise [23, 25]. Bias is defined as the inverse of accuracy: It measures the average difference between the estimation and the *true* value, and it is annotator dependent. An example of bias is demonstrated in Fig. 1 in the context of Electrocardiogram labelling. Recently, Warby *et al.* [20] studied how to combine non-expert annotator's labels of sleep spindle location, a special pattern in human electroencephalography, through fusing annotations provided by non-experts. In that work, although naïve majority vote was used to aggregate the labels of the locations, they demonstrated that non-expert annotations were comparable to those provided by the experts (i.e. the by-subject spindle density correlation was 0.815). Our proposed framework, in contrast, is a statistical

approach that models the precision and bias of each annotator, which we hypothesise would provide a superior estimation of the ground truth as determined by a collection of experts.

In contrast to previous works, this article proposes a Bayesian framework for aggregating multiple continuous-valued annotations in medical data labelling, which takes into account the precision and bias of the individual annotators. Moreover, we propose a generalised form which can be extended to incorporate contextual features of the physiological signal, so that we can adjust the weighting of each label based on the estimated bias and variance of the individual for different types of signal. To our knowledge, the proposed model for estimating continuous-valued labels in an unsupervised manner is novel in the medical domain.

2 Materials and Methods

2.1 Bayesian Continuous-valued Label Aggregator (BCLA)

Suppose that there are N records of physiological time series data labelled by R annotators. Let $\mathbf{D} = \left[\mathbf{x}_i^\mathsf{T}, y_i^{j=1}, \cdots, y_i^{j=R} \right]_{i=1}^N$, where $\mathbf{x}_i$ is a column feature vector for the ith record containing d features (i.e. the design matrix, $\mathbf{X} = [\mathbf{x_1}^\mathsf{T}, ..., \mathbf{x_N}^\mathsf{T}]$), y_i^j corresponds to the annotation provided by the jth annotator for the ith record, and z_i represents the unknown underlying ground truth (the *true* time or duration of an event for example). The graphical representation of the proposed approach – the Bayesian Continuous-valued Label Aggregator (BCLA) – is shown in Fig. 2.

In this model, it is assumed that y_i^j was a noisy version of z_i, with a Gaussian distribution $\mathcal{N}(y_i^j \mid z_i, (\sigma^j)^2)$[1]. Here σ^j is the standard deviation of the jth annotator and represents his variance in annotation around z_i. Furthermore, the bias of each annotator can be modelled as an additional term, ϕ^j. The probability of estimating y_i^j can be written as:

$$\mathrm{P}[y_i^j \mid z_i, (\sigma^j)^2] = \mathcal{N}(y_i^j \mid z_i + \phi^j, 1/\lambda^j). \tag{1}$$

where $(\sigma^j)^2$ is replaced with $1/\lambda^j$. λ^j is the precision of the jth annotator, defined as the estimated inverse-variance of annotator j. Note that λ^j and ϕ^j are considered to be constants for the jth annotator, i.e. all anno-

---

[1] The motivation for this model comes from the Central Limit Theorem. Given the assumption that the annotators are independent and identically distributed, their labels will converge to a Gaussian distribution. In the absence of prior knowledge, this assumption allows for a robust and generalizable model for the given data.

tators are assumed to have consistent but usually different performances throughout records. Furthermore, it is assumed that the probability of a given bias of annotator j, ϕ^j, is drawn from a Gaussian distribution with mean μ_ϕ and variance $1/\alpha_\phi$, is given by:

$$\mathrm{P}[\phi^j \mid \mu_\phi, \alpha_\phi] = \mathcal{N}(\phi^j \mid \mu_\phi, 1/\alpha_\phi). \tag{2}$$

Although the biases of the annotators might be derived from other distributions, they are likely to be data set dependent. In the absence of any knowledge of the underlying distribution of biases, they are assumed to be drawn from a Gaussian distribution. Furthermore, the ground truth, z_i, can be assumed to be drawn from a Gaussian distribution with mean a and variance $1/b$. The probability of z_i is defined as follows:

$$\mathrm{P}[z_i \mid a, b] = \mathcal{N}(z_i \mid a, 1/b), \tag{3}$$

where a can be expressed as a linear regression function $f(\mathbf{w}, \mathbf{x})$ with an intercept, and $\mathbf{w}$ being the coefficients of the regression [16, 26]. The intercept models the overall offset predicted in the regression, which is different from the annotator specific bias in the proposed model. Under the assumption that records are independent, the likelihood of the parameter $\theta = \{\mathbf{w}, \lambda, \phi, \alpha_\phi, b, z_i\}$ for a given data set $\mathbf{D}$ can be formulated as:

$$\mathrm{P}[\mathbf{D} \mid \theta] = \prod_{i=1}^{N} \mathrm{P}[y_i^1, \cdots, y_i^R \mid \mathbf{x}_i, \theta]. \tag{4}$$

It is assumed that $y_i^1, \cdots, y_i^R$ are conditionally independent given the feature $\mathbf{x}_i$ (i.e. each annotator works independently to provide annotations). This may or may not be necessarily true, especially in cases where the annotations are generated by algorithms, some of which may be variations of the same approach. Nevertheless, this assumption was made to simplify the model and subsequent derivation of the likelihood. The likelihood of the parameter θ for a given data set $\mathbf{D}$ can be written using the Bayes' theorem as (see detailed description in Fig. 2):

$$\mathrm{P}[\theta \mid \mathbf{D}] \propto \mathrm{P}[\mathbf{D} \mid \theta] \cdot \mathrm{P}[\theta]$$

$$= \Gamma(\alpha_\phi \mid k_\alpha, \vartheta_\alpha) [\prod_{j=1}^{R} \mathcal{N}(\phi^j \mid \mu_\phi, 1/\alpha_\phi) \Gamma(\lambda^j \mid k_\lambda, \vartheta_\lambda)]$$

$$\Gamma(b \mid k_b, \vartheta_b) [\prod_{i=1}^{N} \mathcal{N}(z_i \mid a, 1/b) \prod_{j=1}^{R} \mathcal{N}(y_i^j \mid z_i + \phi^j, 1/\lambda^j)]. \tag{5}$$

where Γ denotes a Gamma distribution and can be defined as $\Gamma(z \mid k, \vartheta) = \frac{1}{\Gamma(k)\vartheta^k} z^{k-1} exp(-\frac{z}{\vartheta})$, where k is the shape of the distribution and ϑ is the scale of the distribution. Gamma distribution is commonly used to model positive continuous values. It is therefore assumed that precision values, such as b, λ^j, and α_ϕ were drawn from a Gamma distribution, with parameters k_b, ϑ_b, k_λ, ϑ_λ, and k_α, ϑ_α respectively.

2.2 The Maximum a posteriori approach

The estimation of θ can be solved using the maximum a posteriori (MAP) approach, which maximises the log-likelihood of the parameters, i.e. $\underset{\theta}{\operatorname{argmax}}\{\log P[\theta \mid \mathbf{D}]\}$. The log-likelihood can be rewritten as:

$$
\begin{aligned}
\log P[\theta \mid \mathbf{D}] = &-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{R}[\log(\frac{2\pi}{\lambda^j}) + (y_i^j - \phi^j - z_i)^2\lambda^j] \\
&-\frac{1}{2}\sum_{j=1}^{R}[\log(\frac{2\pi}{\alpha_\phi}) + (\phi^j - \mu_\phi)^2\alpha_\phi] \\
&-\frac{1}{2}\sum_{i=1}^{N}[\log(\frac{2\pi}{b}) + (z_i - \mathbf{x}_i^\mathsf{T}\mathbf{w})^2 b] \\
&+[(k_\lambda - 1)\log\lambda^j - \log(\Gamma(k_\lambda)\vartheta_\lambda^{(k_\lambda)}) - \frac{\lambda^j}{\vartheta_\lambda}] \\
&+[(k_\alpha - 1)\log\alpha_\phi - \log(\Gamma(k_\alpha)\vartheta_\alpha^{(k_\alpha)}) - \frac{\alpha_\phi}{\vartheta_\alpha}] \\
&+[(k_b - 1)\log b - \log(\Gamma(k_b)\vartheta_b^{(k_b)}) - \frac{b}{\vartheta_b}].
\end{aligned}
$$
(6)

The parameters in θ can be derived by equating the gradient of the log-likelihood to zero respectively as follows:

$$
\frac{1}{\lambda^j} = \frac{1}{N + 2(k_\lambda - 1)}[\sum_{i=1}^{N}(y_i^j - \phi^j - z_i)^2 + \frac{2}{\vartheta_\lambda}].
$$
(7)

$$
\mathbf{w} = (\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\mathsf{T})^{-1}\sum_{i=1}^{N}\mathbf{x}_i z_i.
$$
(8)

$$
\phi^j = \frac{1}{N + \frac{\alpha_\phi}{\lambda^j}}[\sum_{i=1}^{N}(y_i^j - z_i) + \mu_\phi(\frac{\alpha_\phi}{\lambda^j})].
$$
(9)

$$
\frac{1}{\alpha_\phi} = \frac{1}{R + 2(k_\alpha - 1)}[\sum_{j=1}^{R}(\phi^j - \mu_\phi)^2 + \frac{2}{\vartheta_\alpha}].
$$
(10)

$$
z_i = \frac{\sum_{j=1}^{R}[(y_i^j - \phi^j)\lambda^j] + (\mathbf{x}_i^\mathsf{T}\mathbf{w})b}{\sum_{j=1}^{R}\lambda^j + b}.
$$
(11)

$$
\frac{1}{b} = \frac{1}{N + 2(k_b - 1)}[\sum_{i=1}^{N}(z_i - \mathbf{x}_i^\mathsf{T}\mathbf{w})^2 + \frac{2}{\vartheta_b}].
$$
(12)

This MAP problem can be solved using the EM algorithm in a two-step iterative process:

i) The E-step estimates the expected *true* annotations for all records, $\hat{\mathbf{z}}$, as a weighted sum of the provided annotations, and can be estimated using equation (11).

ii) The M-step is based on the current estimation of $\hat{\mathbf{z}}$ and given the data set $\mathbf{D}$. The model parameters, $\mathbf{w}$, ϕ, α_ϕ, b, and λ can be updated using equations (8), (9), (10), (12), and (7) accordingly in a sequential order until convergence, which is now described.

2.3 Convergence criteria for the MAP-EM approach

When solving a MAP-EM algorithm one may encounter a convergence issue, particularly when estimating a large number of parameters. The estimation of the precision may approach to infinity because the inferred annotations favour the annotator with the highest precision in each EM update step while maximising the likelihood. Instead of incorporating an additional parameter for the regularisation penalty that increases the complexity of the mode, the generalized extreme value distribution (GEVD) can be used to model the maxima of the precision distribution, denoted as λ_m, in order to restrict the upper bound of the precision values and guarantee a convergence in the MAP algorithm. The probability density function of the GEVD for λ_m can be expressed as:

$$\mathrm{P}(\lambda_m \mid k, \mu, \vartheta) = \exp\{-[1 + k\frac{(\lambda_m - \mu)}{\vartheta}]^{-\frac{1}{k}}\}\frac{1}{\vartheta}[1 + k\frac{(\lambda_m - \mu)}{\vartheta}]^{(-1 - \frac{1}{k})}, \tag{13}$$

where k is the shape parameter, ϑ is the scale parameter, and μ is the location parameter. These parameters can be derived by fitting a GEVD to the maximum values drawn randomly from the *prior* distribution of the precision, $\Gamma(\lambda \mid k_\lambda, \vartheta_\lambda)$. An upper bound of the maximum precision value can then be obtained by estimating the $99th$ quantile of the inverse cumulative distribution function of the GEVD.

2.4 Data description

The electrocardiogram (ECG) is a standard and powerful tool for assessing cardiovascular health as many detrimental heart conditions manifest as abnormalities in the ECG. The QT interval is one particular measure of ECG morphology, and refers to the elapsed time between the onset of ventricular depolarisation (the QRS complex) and the T wave offset (ventricular repolarisation) [4]. Accurate measurement of the QT interval is essential since abnormal intervals indicate a potentially serious but treatable condition, and can be a contraindication for the use of drugs or other interventions [11]. Viskin *et al.* [19] presented the ECGs recorded from two patients with long QT syndrome (LQTS) and from two healthy females to 902 physicians (25 QT experts who had published on the subject, 106 arrhythmia specialists, 329 cardiologists, and 442 noncardiologists) from 12 countries. No other details were given on actual training or intrinsic accuracy of these annotators. For patients with LQTS, 80% of arrhythmia specialists calculated the QTc (the heart rate corrected QT interval) correctly but only 50% of cardiologists and 40% of noncardiologists did so. In the context of QT annotation where baseline wander is frequent,

it was observed that a few annotators consistently over- or under-estimated the QT interval [25]. Other studies have reported significant intra- and inter-observer variability in QT annotations, ranging from 10 to 30ms [3, 8]. It is important to note that experts or non-experts with different levels of training or expertise can have significantly different biases. Naïve approaches to aggregate labels from a group of annotators of unknown expertises could therefore lead to poor results. However, annotators' biases are rarely taken into account when aggregating different labels or opinions in medical labelling tasks.

We hypothesise that incorporating an accurate estimation of each annotator's bias into a model for fusing annotations (as described in sections 2.1 to 2.3) will result in an improved estimate of the ground truth. In order to test this hypothesis we have used two data sets: one simulated data set to ensure an absolute ground truth is available; and one real data set of QT intervals. Although we have chosen to use QT interval data, because of the availability of the numerous annotations, the method we present is more general and can be applied to other continuous-valued annotations..

2.4.1 Simulated data set

To test the reliability of the BCLA as a generative model, a simulated data set was created: a total of 548 simulated records were generated, each has 20 independent annotator, thus providing a total of 10,960 annotations (see Fig. 3). The simulated data set considered that annotators have precision values, λ (i.e. $1/\sqrt{\sigma}$), which were drawn from $\Gamma(4, 0.0003)$, with assumption that the annotations provided by the best performing annotator is ± 15ms away from the ground truth. Annotators' biases were drawn from $\mathcal{N}(10, 25)$, a Gaussian distribution with 10ms mean and a standard deviation ($1/\sqrt{\alpha_\phi}$) of 25ms. The *true* annotation for each record was drawn from $\mathcal{N}(400, 40)$, a Gaussian distribution with a mean, a, of 400ms with a standard deviation ($1/\sqrt{b}$) of 40ms. In addition, it was assumed that α_ϕ was drawn from $\Gamma(3, 0.0005)$, ensuring the mean standard deviation where the biases drawn from is 25ms. The b was drawn from $\Gamma(3, 0.0002)$, ensuring the mean standard deviation where the *true* annotations drawn from is 40ms. The generated 10,960 annotations were then fed into the BCLA model to evaluate its accuracy in estimating the *true* annotation in an unsupervised manner as well as predicting the bias and precision of each annotator.

2.4.2 Real data set

The data were drawn from the QT interval annotations generated by participants in the 2006 PhysioNet/Computing in Cardiology (PCinC) Challenge [15] for labelling QT intervals with reference to Lead II in each of the 548 recordings in the Physikalisch-Technische Bundesanstalt Diagnostic ECG Database (PTBDB) [2]. The records were from 290 subjects (209 men with mean age of 55.5 years and 81 women with mean age of 61.6 years), in which 20% of the subjects were healthy controls. An example of QT interval is demonstrated in Fig. 1(c). The PTBDB database contained records of patients with a variety of ECG morphologies having different QT intervals ranging from 256 to 529 ms. The diagnostic classifications of ECG morphologies mainly included myocardial infarction, heart failure, bundle branch block, and dysrhythmia as stated in Bousseljot and Kreiseler [2].

There were two main categories of annotations: manual and automated (see Table 1). A total of 38,621 annotations were collected and were divided into three divisions: 20 human annotators in Division 1, 48 closed source automated algorithms in Division 2, and 21 open source automated algorithms in Division 3. Division 4 was further created here so as to combine all automated algorithms from Division 2 and 3 in order to provide a larger data set and allow a better estimation of automated QT intervals. The number of annotators per division and averaged number of annotations per record are listed in Table 1. The overall percentage of the annotators in each division with complete annotations (i.e. annotations on all 548 recordings) was: 55% in Division 1, 40% in Division 2, 43% in Division 3, and 45% in Division 4. The competition score for each entry was calculated from the root mean square error (RMSE) between the submitted and the reference QT intervals. The reference annotations were generated from Division 1's entries using a maximum of 15 participants by taking the "median self-centering approach" as reported by the competition organisers as detailed in [24]. The best-performing score for each division is also listed in Table 1. Furthermore, the majority of the QT annotations of each 2-minute record occurred within the first 5 seconds of the ECG recordings. The best scores in the first 5-second segment were similar to those of the 2-minute segment (denoted by $\star$ in Table 1). To reduce any possible inter-beat variations, only the annotations within the first 5-second segment of each record were chosen to ensure that all annotators had approximately labelled the same region of a record with similar QT morphologies. Therefore, the motivation for choosing the first 5-second segment of each record was to consider a short segment where the QT interval is not changing dramatically (with respect to a particular beat an annotator chose), while retaining the highest num-

ber of annotations. Those that fell outside this segment were considered to be missing information and discarded in the process of the QT estimation.

As the manual entry (i.e. Division 1) was used to generate the reference annotations, we therefore focused on the analysis of the automated entry (i.e. Division 2, 3, and 4). In terms of parameter setting (see Table 2), annotator specific precision was drawn from $\Gamma(k_\lambda, \vartheta_\lambda)$, with assumption that the annotations provided by the best performing algorithm is ± 5ms away from the reference. Annotators' biases were considered to be drawn from $\mathcal{N}(\mu_\phi, 1/\sqrt{\alpha_\phi})$, and α_ϕ was modelled by $\Gamma(k_\alpha, \vartheta_\alpha)$, assuming that the automated annotations tend over-estimate manual annotations as described in previous studies [1, 5, 10]. The *true* QT interval for each record was assumed to be drawn from $\mathcal{N}(a, 1/\sqrt{b})$, where b was modelled by $\Gamma(k_b, \vartheta_b)$ [4, 9, 12]. Instead of assuming the mean (i.e. a) of the underlying ground truth to be a fixed scalar, we updated it using a linear regression function, $f(\mathbf{w}, \mathbf{x})$, where the coefficients, $\mathbf{w}$, were estimated using equation (8). An intercept was included in $f(\mathbf{w}, \mathbf{x})$ to model the overall offset predicted in f, and no particular features were considered in this case (i.e. $x_i = 1$) as we were solely interested in the performance of the model.

2.5 Methodology of validation and comparison

The BCLA inferred precision of individual algorithms was compared with those estimated using the EM algorithm proposed by Raykar *et al.* [16] (denoted as EM-R) as it served as one of the benchmarking algorithms. Furthermore, the mean and standard deviation ($\mu \pm \sigma_\mu$ms) of 100 bootstrapped (i.e. random sampling with replacement) samples across records from the BCLA model were compared with the best algorithm (i.e. the algorithm with highest precision after correction of the bias offset), EM-R, and the traditional naïve mean and median voting approaches in both simulated and real data sets. The mean absolute error (MAE) of the annotations was also calculated as it provides interpretation of the difference between the estimated and the reference annotations (with a resolution of 1ms). A two-sided Wilcoxon rank sum test ($p < 0.0001$) was applied to the 100 bootstrapped RMSEs and MAEs, to provide a comparison for the BCLA and EM-R versus other methodologies. In assessing the performance of the BCLA as a function of the number of annotators, a random number of annotators was selected 100 times. This was repeated with the annotator numbers varied from three to the maximum number of annotators in the division. The minimum number of annotators was chosen to be three to allow for obtaining

results from the median voting approach. The $\mu \pm \sigma_\mu$ ms of the RMSE of the BCLA, the EM-R, the mean, and the median were calculated and compared.

3 Results

The convergence of the BCLA model is guaranteed by providing a threshold using the GEVD as a stopping criteria (see Eqn (13)). In the real data set, the upper bound of the precision derived from the GEVD was 0.04, which was based on the assumption that the best performing annotator is ± 5ms away from the reference. The number of iteration is dependent on the number of records and the number of annotations. To illustrate the practical utility of our model, it took 7.55 seconds for the BCLA to perform 5,000 iterations when considering a total of 20,712 annotations (Division 2) using MATLAB R2011a on a 2.2GHz Intel(R) i7-2670QM processor. Approximately 2,500 iterations were required to stabilise all the parameters.

3.1 Simulated data set

Fig. 4(a) shows an example of the inferred results estimated using the EM-R and the BCLA. As the EM-R algorithm modelled jointly the precision (i.e. $1/(\sigma)^2$) of each annotator and the noise of underlying ground truth, its estimated σ cannot represent the real precision of each annotator. Furthermore, EM-R algorithm does not consider the bias of each annotator, and we observe that its estimated values of σ were well above the line of identity, indicating a consistent over-estimation. In contrast, the BCLA inferred results of σ lie closely to the line of identity in the plot, indicating that the BCLA model can provide a reliable estimation of the *true* precision in the simulated results. In addition to precision, the BCLA modelled the bias of each annotator and the results are provided in Fig. 4(b): the estimated biases are very close to the *true* biases. Although not all the estimated precisions and biases of each annotator were identical to the simulated values, the BCLA model inferred annotations without any prior knowledge of who the best annotator was in an unsupervised manner.

In order to compare the accuracy of the inferred labels using the BCLA model, the simulated 548 annotations were bootstrapped 100 times. Each time a RMSE and MAE were generated and compared to the best annotator, mean, EM-R, and median voting strategies. The results are shown in Table 3. The RMSE and MAE results

show that BCLA inferred labels significantly outperformed the mean, median, EM-R, and best annotator when compared with the simulated *true* annotations.

3.2 Real data set

Fig. 5 (a) to (f) show the inferred precision and bias results estimated using EM-R and BCLA for different automated divisions. As mentioned previously, the EM-R algorithm does not directly model the precision (i.e. $1/(\sigma)^2$) of each annotator; its estimated σ of each annotator produces an offset from the values provided by the reference annotations. In contrast, the BCLA inferred σ results lie much closer to the line of identity in the Fig. 5 (a), (c), and (e), indicating that the BCLA model can provide a reliable estimation of the *true* precision of each annotator. In addition, the BCLA modelled the bias of each annotator accurately (see Fig. 5 (b), (d), and (f)). Although automated annotator 3 and 15 were predicted by the BCLA to have lower bias values than those provided by the reference, they are considered to be outliers due to the assumption made in our model: annotators' biases were drawn from a Gaussian distribution with 10ms mean and 25ms standard deviation. As Fig. 5 (g) shows, the biases of annotator 3 and 15 lie outside the 95% of the area (i.e. $\pm 1.96\sigma$ of the mean under the normal distribution) predicted by the BCLA. In the case of annotator 7, its precision was underestimated (see Fig. 5(c) and (e)), which also affected the BCLA's estimation of its bias value. It was observed that only 3.47% of records were annotated by annotator 7, making it harder for the BCLA to provide a reliable estimation of its precision and bias values. In the evaluation of the inferred labels, the 548 records were bootstrapped 100 times, the RMSEs and MAEs of the BCLA model were generated and compared to the best annotator, mean, EM-R, and median voting approaches for the given reference. The results are displayed in Table 4:

for Division 2 using 48 algorithms, the BCLA achieved a RMSE of 12.57$\pm$0.67ms, which significantly outperformed other approaches and provides an improvement of 16.48% over the next best approach (EM-R with RMSE of 15.05$\pm$0.49ms); in the closed source entry Division 3 using 21 algorithms, the BCLA again exhibited a superior performance over the other methods with a RMSE of 13.90$\pm$0.84, and a 19.48% improved error rate over the next best method (RMSE of 17.25$\pm$2.33ms). When considering all automated entries (Division 4), the BCLA provided an even more accurate performance than on the other two data sets (Division 2 and 3) as well as over other methods tested with a RMSE of 11.78$\pm$0.63ms.

A further evaluation of the accuracies in terms of RMSE were made as a function of the number of annotators (see Fig. 6). The results were generated by sub-sampling annotators with no replacement 100 times. The benchmarking algorithm, EM-R outperformed mean and median approaches initially but then underperformed when compared to the median approach after 43 algorithms are used. The BCLA model outperformed the other methods being tested with any number of annotators considered. In practice, it is rare to have more than three to five independent algorithms for estimating a label or predicting an event. In the case where only three automated algorithms were randomly selected, the BCLA had on average 9.02%, 19.82%, and 24.56% improvement over the EM-R, median and mean voting approaches respectively.

Although the lowest BCLA RMSE (11.78$\pm$0.63ms) in the automated entry is larger than the best-performing human annotator in the Challenge (RMSE = 6.65ms), there were only two other human annotators who achieved a score below 10ms. Furthermore, as the annotations of automated algorithms were independently determined from the reference, whereas the reference includes the best human annotators, it is unsurprising that a combination of the automated algorithms would have worse performance.

4 Discussion

In this article, a novel model, Bayesian Continuous-valued Label Aggregator, was proposed to infer the ground truth of continuous-valued labels where accurate and consistent expert annotations are not available. As a proof-of-concept, the BCLA was applied to the QT interval estimation from the ECG using labels from the 2006 PhysioNet/Computing in Cardiology Challenge database, and it was compared to the mean, median, and a previously proposed Expectation Maximization label aggregation methods (i.e. EM-R). While accurately predicting each labelling participants bias and precision, the root-mean-square error of the BCLA algorithm was significantly outperformed the best Challenge entry as well as the EM-R, mean, and median voting strategies. There are two key contributions in our approach: i) the BCLA provides an estimation of *continuous-valued annotations* which is valuable for time-series related data as well as duration of events for physiological data; ii) It introduces a unified framework for combining *continuous-valued annotations* to infer the underlying ground truth, while *jointly modelling annotators' biases and precisions*. The BCLA operates in an unsupervised Bayesian learning framework; no reference data were used to train the model parameters and a separate training and validation test sets were

not required. Combining more experienced annotators would therefore provide a better estimation of the inferred ground truth. Importantly though, the BCLA does *guarantee a performance better than the best annotator without any prior knowledge of who or what is the best annotator*.

Novel contextual features were introduced in our previous study [26] which allowed an algorithm to learn how varying physiological and noise conditions affect each annotator's ability to accurately label medical data. The inferred result was shown to provide an improved 'gold standard' for medical annotation tasks even when the ground truth is not available. As the next step, if we incorporate the context into the weighting of annotators, the BCLA is expected to have an even larger impact for noisy data sets or annotators with a variety of specialisations or skill levels. The current model assumed consistent performance of each annotator throughout the records: i.e. that is his/her performance is time-invariant. Although this might not be true over an extended period of time where an annotators performance might improve through learning, or their performance might drop due to inattention or fatigue, the nature of the data sets being considered in this work are such that we can assume that performance across records is approximately consistent for each annotator. Future work will include modelling the performance of each annotator varying across records and through time to provide a more reliable estimation of the aggregated ground truth for data sets in which intra-annotator performance is highly variant.

Our model of the annotators currently does not factor in the possible dependency/correlation between individual annotators, which might not be the case for automated algorithms. Incorporating a correlation measure into the annotator's model could possibly allow for a better aggregation of the inferred ground truth. Annotators who are considered to be anomalous (i.e. highly correlated but have large variances and biases) should be penalised with lower weights; expert annotators (i.e. highly correlated but have small variances and biases) should be favourably voted in the model. Finally, combining annotations derived from reliable experts using the BCLA model could potentially lead to improved training for supervised labelling approaches.

References

1. Andrew, W., Michael, V., Jeff, D., Nair, G.M., Plater-Zyberk, C., Griffith, L., Ma, J., Zachos, C., Sivilotti, M.L.: Variability of QT Interval Measurements in OpioidDependent Patients on Methadone. CJAM **2**, 10–16 (2014)

2. Bousseljot R Kreiseler D, S.A.: Nutzung der EKG-Signaldatenbank CARDIODAT der PTB uber das Internet. Biomed Tech **40(1)**, 317–318 (1995)

3. Christov, I., Dotsinsky, I., Simova, I., Prokopova, R., Trendafilova, E., Naydenov, S.: Dataset of manually measured QT intervals in the electrocardiogram. Biomed Eng Online **5**, 31 (2006)

4. Clifford, G.D., Azuaje, F., McSharry, P.E.: Advanced Methods and Tools for ECG Analysis. Engineering in Medicine and Biology. Artech House, Norwood, MA, USA (2006)

5. Couderc, J.P., Garnett, C., Li, M., Handzel, R., McNitt, S., Xia, X., Polonsky, S., Zareba, W.: Highly Automated QT Measurement Techniques in 7 Thorough QT Studies Implemented under ICH E14 Guidelines. Ann Noninvasive Electrocardiol **16**(1), 13–24 (2011)

6. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the EM algorithm. Appl Stat-J Roy St C **28**(1), 20–28 (1979)

7. Dekel, O., Shamir, O.: Good learners for evil teachers. In: Proc 26th Annu ICML, ICML '09, pp. 233–240. ACM (2009)

8. Ehlert, F.A., Goldberger, J.J., Rosenthal, J.E., Kadish, A.H.: Relation between QT and RR intervals during exercise testing in atrial fibrillation. Am J Cardiol **70**(3), 332–338 (1992)

9. Goldenberg, I., Moss, A.J., Zareba, W., et al.: QT interval: how to measure it and what is "normal". J Cardiovasc Electrophysiol **17**(3), 333–336 (2006)

10. Hughes, N.P.: Probabalistic Models for Automated ECG Interval Analysis. Ph.D. thesis, University of Oxford (2006)

11. International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use: Guidance for Industry E14: Clinical Evaluation of QT/ QTc Interval Prolongation and Proarrhythmic Potential for Non- Antiarrhythmic Drugs (2014)

12. Malik, M., Frbom, P., Batchvarov, V., Hnatkova, K., Camm, A.J.: Relation between QT and RR intervals is highly individual among healthy subjects: implications for heart rate correction of the QT interval. Heart **87**(3), 220–228 (2002)

13. Metlay, J.P., Kapoor, W.N., Fine, M.J.: Does this patient have community-acquired pneumonia?: Diagnosing pneumonia by history and physical examination. J Am Coll Cardiol **278**(17), 1440–1445 (1997)

14. Molinari, F., Gentile, L., Manicone, P., Ursini, R., Raffaelli, L., Stefanetti, M., D'Addona, A., Pirronti, T., Bonomo, L.: Interobserver variability of dynamic MR imaging of the temporomandibular joint. La Radiologia Medica **116**(8), 1303–1312 (2011)

15. Moody, G.B., Koch, H., Steinhoff, U.: The PhysioNet/ Computers in Cardiology Challenge 2006: QT interval measurement. In: Comput Cardiol, pp. 313 –316 (2006)

16. Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. JMLR pp. 1297–1322 (2010)

17. Salerno, S.M., Alguire, P.C., Waxman, H.S.: Competency in interpretation of 12-lead electrocardiograms: a summary and appraisal of published evidence. Ann Intern Med **138**(9), 751–760 (2003)

18. Valizadegan, H., Nguyen, Q., Hauskrecht, M.: Learning Medical Diagnosis Models from Multiple Experts. In: AMIA Annu Symp Proc, pp. 921–930. AMIA (2012)

19. Viskin, S., Rosovski, U., Sands, A.J., Chen, E., Kistler, P.M., Kalman, J.M., Chavez, L.R., Torres, P.I., CruzF, F.E., Centurion, O.A., Fujiki, A., Maury, P., Chen, X., Krahn, A.D., Roithinger, F., Zhang, L., Vincent, G.M., Zeltser, D.: Inaccurate electrocardiographic interpretation of long QT: the majority of physicians cannot recognize a long QT when they see one. Heart Rhythm **2**, 569–574 (2005)

20. Warby, S.C., Wendt, S.L., Welinder, P., Munk, E.G., Carrillo, O., Sorensen, H.B., Jennum, P., Peppard, P.E., Perona, P., Mignot, E.: Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. Nat methods **11**(4), 385–392 (2014)

21. Warfield, S.K., Zou, K.H., Wells, W.M.: Validation of image segmentation by estimating rater bias and variance. Philos T R Soc A **366**, 2361–2375 (2008)

22. Welinder, P., Branson, S., Perona, P., Belongie, S.J.: The multidimensional wisdom of crowds. In: Adv Neural Inf Process Syst 23, pp. 2424–2432 (2010)

23. Welinder, P., Perona, P.: Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In: IEEE CVPRW, pp. 25–32 (2010)

24. Willems, J., Arnaud, P., van Bemmel, J., Bourdillon, P., Brohet, C., Dalla Volta, S., Andersen, J., Degani, R., Denis, B., Demeester, M., et al.: Assessment of the performance of electrocardiographic computer programs

with the use of a reference data base. Circulation **71(3)**, 523 – 534 (1985)

25. Zhu, T., Behar, J., Papastylianou, T., Clifford, G.D.: Crowdlabel: A Crowd-sourcing Platform for Electrophysiology. In: Comput Cardiol, vol. 41 (2014)

26. Zhu, T., Johnson, A.E., Behar, J., Clifford, G.D.: Crowd-Sourced Annotation of ECG Signals Using Contextual Information. Ann Biomed Eng **42**(4), 871–884 (2014). DOI 10.1007/s10439-013-0964-6
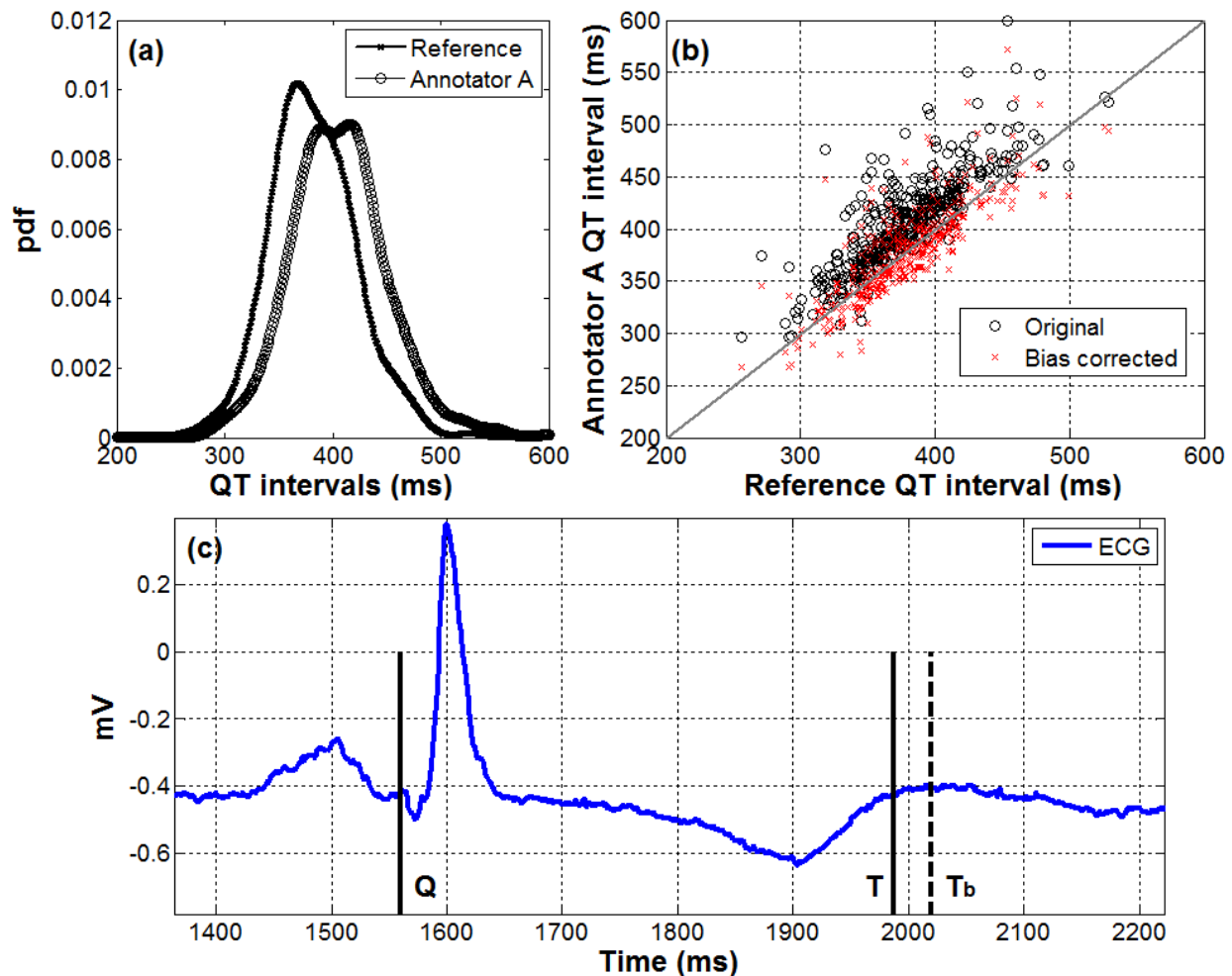
Fig. 1 An example of bias in the context of Electrocardiogram (ECG) QT interval labelling. (a) The probability density function of the QT intervals for the reference (supplied by the human experts) annotation and annotator A (such as an automated algorithm). A plot of QT intervals across different recordings: the diagonal (grey) line indicates a perfect match of QT intervals between the reference and annotator A; the 'o' indicates the original QT intervals provided by annotator A; the 'x' indicates the bias corrected QT intervals of annotator A, which fits closely to the diagonal line. (c) An example of bias that occurs in an ECG record for labelling QT interval. The reference QT interval on a single beat starts at the beginning of the Q wave and ends at the end of the T wave (denoted as Q and T), and the biased trend from annotator A is demonstrated as T_b.
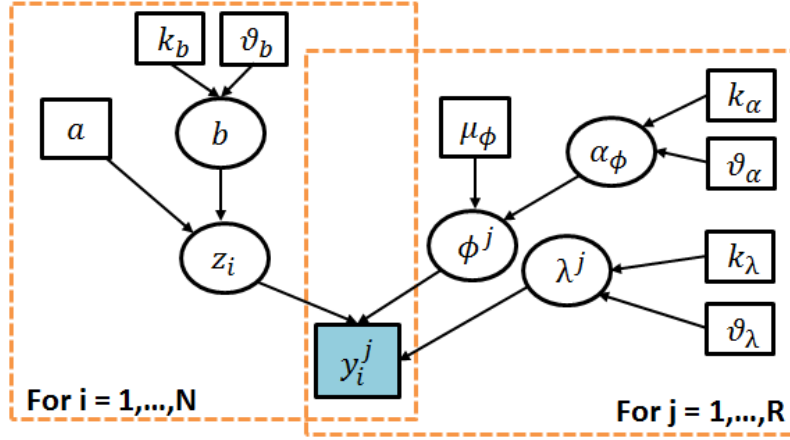
Fig. 2 Graphical representation of the BCLA model: y_i^j corresponds to the annotation provided by the jth annotator for the ith record, and it is modelled by the z_i (the unknown underlying ground truth), the ϕ^j (bias), and the λ^j (precision). Furthermore, z_i is drawn from a Gaussian distribution with parameters mean a and variance $1/b$, where a can be a function of feature vector $\mathbf{x}_i$. ϕ^j is modelled from a Gaussian distribution with mean μ_ϕ and variance $1/\alpha_\phi$. The b, λ^j, and α_ϕ are drawn from a Gamma distribution (denoted as Γ) with parameters k_b, ϑ_b, k_λ, ϑ_λ, and k_α, ϑ_α respectively.
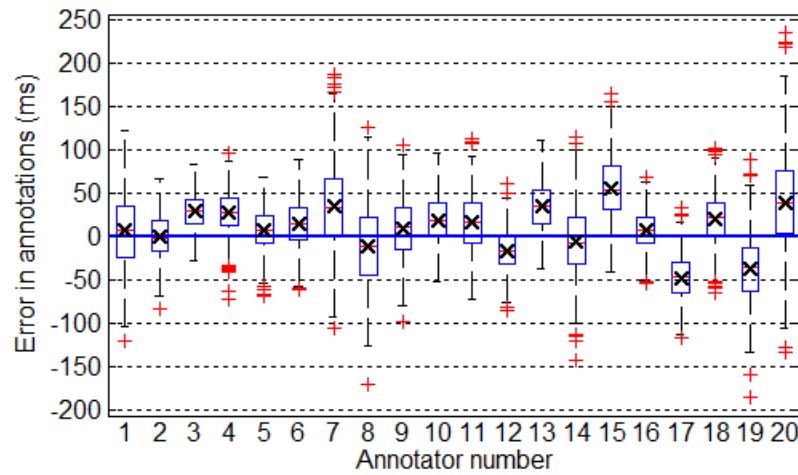
Fig. 3 The box plot of the error between the generated and *true* annotations for each of the 20 simulated annotator. The black 'x' indicates the bias of each annotators. The span of each box represents the precision of the annotations (rather than the interquartile range) over all annotations for each annotator.

Table 1 Performance by competition entrants on the first 5-second ECG segment for each division of the 2006 PCinC Challenge.

	Manual annotators	Automated algorithms		
	Division 1	Division 2	Division 3	Division 4
Number of annotators	20	48	21	69
Average annotations per record	18 (18⋆)	39 (41⋆)	15 (21⋆)	54 (62⋆)
RMSE score (ms)	6.65 (6.67⋆)	16.36 (16.34⋆)	17.46 (17.33⋆)	16.36 (16.34⋆)
Interquartile range of score (ms)	30.40	35.77	128.00	57.00

Note: The annotator/algorithm having the lowest RMSE over the 5-second segment was selected to represent the best score. The results with ⋆ were published in the Challenge for a 2-minute segment.
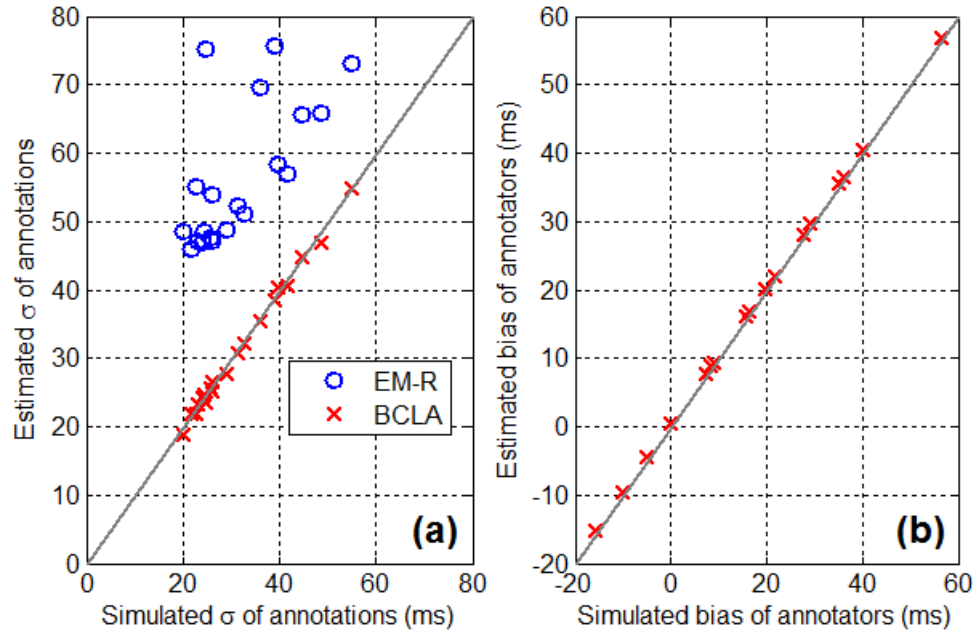
Fig. 4 A comparison of the simulated and inferred σ in (a) and bias in (b) of each annotator in the simulated data set. The precision can be estimated by taking $1/(\sigma)^2$. The diagonal (grey) line indicates a perfect match between simulated and estimated results. Note that EM-R significantly over-estimates the σ in all simulations.

Table 2 The parameters of the BCLA and their values for modelling the 2006 PCinC data set.

Symbol	Definition	Value
k_b	shape of Gamma distribution for b	3 ‡
ϑ_b	scale of Gamma distribution for b	0.0002 ‡
μ_ϕ	mean of the bias distribution	10 †
k_α	shape of Gamma distribution for α_ϕ	3 †
ϑ_α	scale of Gamma distribution for α_ϕ	0.0005 †
k_λ	shape of Gamma distribution for λ	4*
ϑ_λ	scale of Gamma distribution for λ	0.003*

Note: b is the precision parameter for the model of the ground truth. α_ϕ is the precision parameter for the model of the bias. λ refers to annotators' precision values. The values with * are determined with the assumption that the annotations provided by the best performing algorithm is ± 5ms away from the reference. The values with † are derived from [1, 5, 10]. The values with ‡ are derived from [4, 9, 12].

Table 3 The RMSEs and the MAEs of the inferred labels using different strategies in the simulated data set.

	Best Annotator	Median	Mean	EM-R	BCLA
RMSE (ms)	34.91±0.74*	18.84±0.38*	13.11±0.31*	14.21±0.36	6.44±0.34*†
MAE (ms)	30.15±0.72*	12.60±0.36	11.26±0.30*	12.64±0.36	5.14±0.30*†

Results significantly different from others ($p < 0.0001$) as shown in † for the BCLA model and * (columns 2 to 4, and 6 only) for the EM-R.
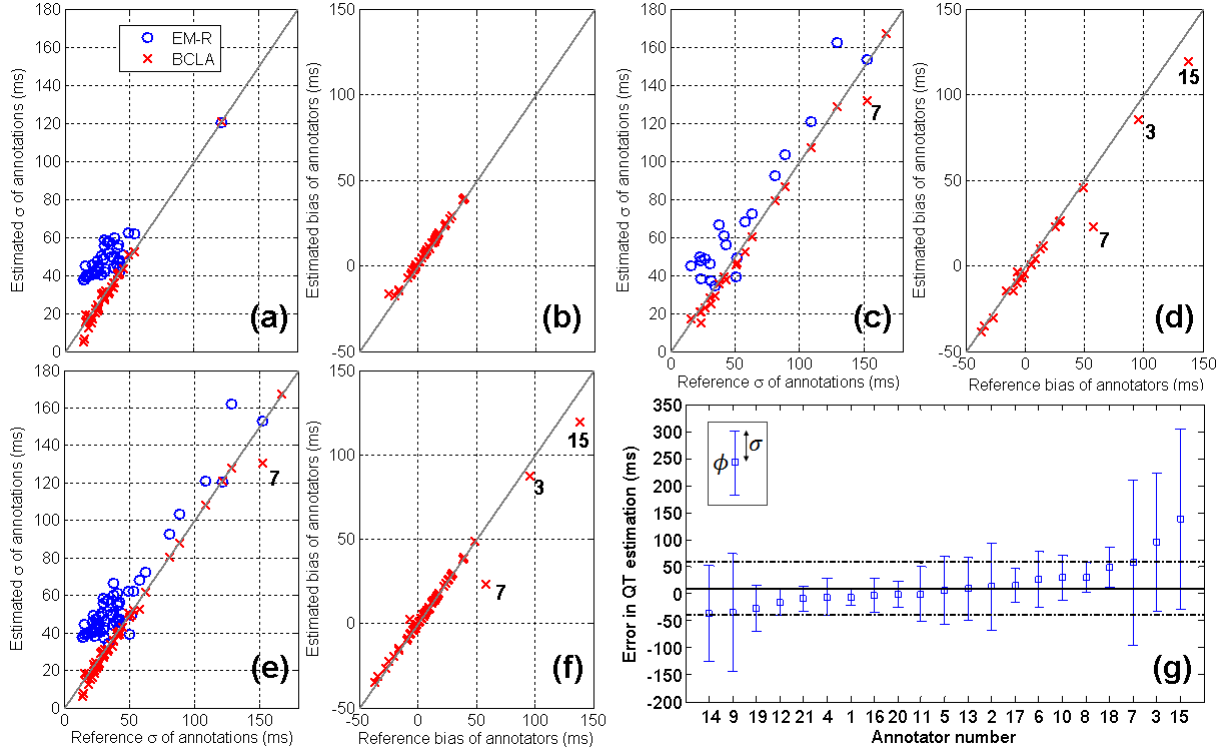
Fig. 5 A comparison of the 2006 PCinC Challenge reference and inferred σ and bias of each annotator using the reference provided for division 2 in (a) and (b), division 3 in (c) and (d), and division 4 in (e) and (f) respectively. The precision can be estimated by taking $1/(\sigma)^2$. The leading diagonal line of each plot indicates a perfect matched between the Challenge reference and the estimated results. The mean (i.e. bias), ϕ, and σ of the difference in annotations for Division 3 are shown in (g). The annotators were ranked based on their bias values. The solid line indicates the mean of the biases whereas the dotted lines indicate 1.96σ of the mean assumed in the BCLA. Note the annotator 3, 7, and 15 are labelled in the corresponding plots.

Table 4 The RMSEs and the MAEs of the inferred labels using different voting approaches in the 2006 PCinC data set.

	RMSE (ms)				
Div	**Best Annotator**	**Median**	**Mean**	**EM-R**	**BCLA**
2	15.43±0.73*	15.29±0.58	16.17±0.54*	15.05±0.49	12.57±0.67*†
3	17.25±2.33*	19.16±0.88	30.46±1.57*	18.92±0.82	13.90±0.84*†
4	15.37±2.13*	14.43±0.57*	17.61±0.55*	14.76±0.52	11.78±0.63*†
	MAE (ms)				
Div	**Best Annotator**	**Median**	**Mean**	**EM-R**	**BCLA**
2	10.85±0.58*	11.76±0.42	12.61±0.43*	11.81±0.40	9.29±0.45*†
3	11.61±3.03*	14.04±0.55	22.89±0.96*	14.12±0.60	10.28±0.67*†
4	11.17±2.32*	11.21±0.40*	14.16±0.43*	11.49±0.41	8.56±0.42*†

Results significantly different from others ($p < 0.0001$) as shown in † for the BCLA model and * (columns 2 to 4, and 6 only) for the EM-R.
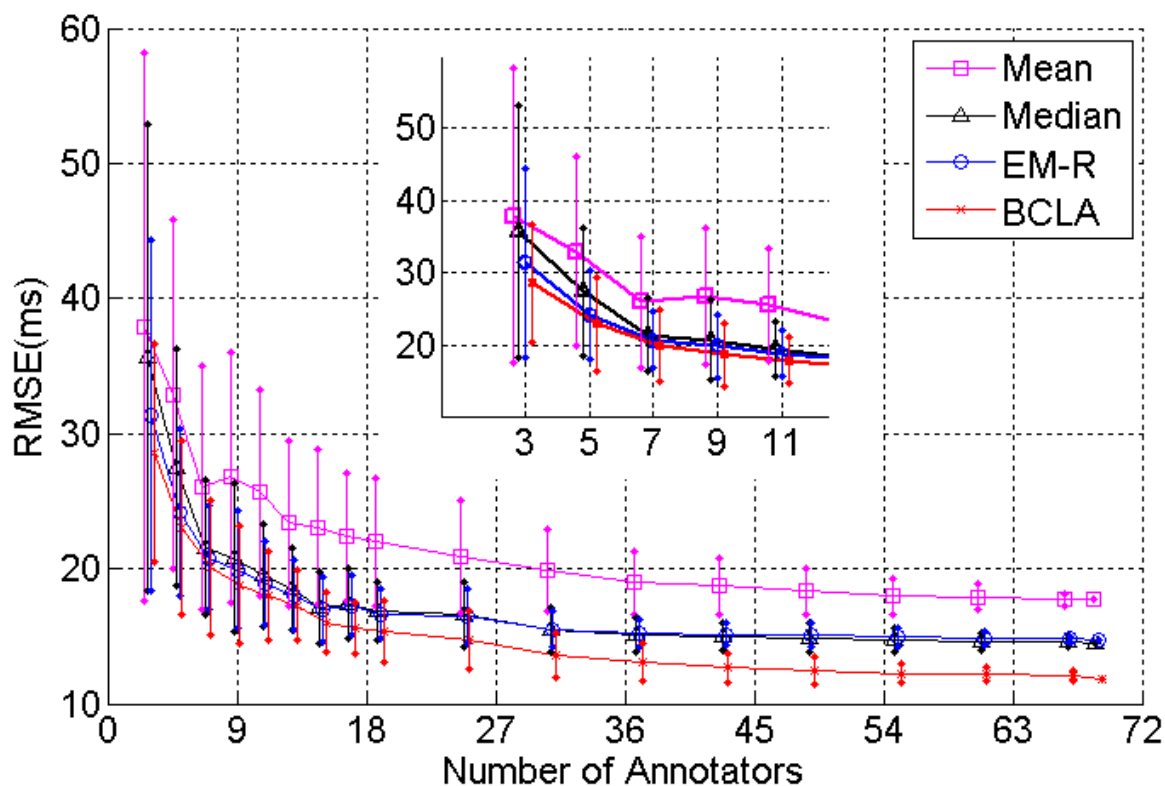
Fig. 6 The mean and standard deviation of the RMSE results as a function of the number of annotators for Division 4 when using the BCLA, EM-R, median, and mean voting approaches. Inset: A close-up of the RMSE results when using 11 annotators or less.